Overview of the 1st International Workshop on Interactive Video Search and Exploration

Luca Rossetto
Dublin City University
Dublin, Ireland
luca.rossetto@dcu.ie

George Awad
NIST
Gaithersburg, MD, USA
george.awad@nist.gov

Werner Bailer
Joanneum Research
Graz, Austria
werner.bailer@joanneum.at

Cathal Gurrin
Dublin City University
Dublin, Ireland
cathal.gurrin@dcu.ie

Björn Þór Jónsson
Reykjavik University
Reykjavík, Iceland
bjorn@ru.is

Jakub Lokoč
Charles University
Prague, Czech Republic
jakub.lokoc@matfyz.cuni.cz

Stevan Rudinac
University of Amsterdam
Amsterdam, The Netherlands
s.rudinac@uva.nl

Klaus Schoeffmann
Klagenfurt University
Klagenfurt, Austria
ks@itec.aau.at

Abstract

The inaugural 1st International Workshop on Interactive Video Search and Exploration (IViSE), held at CVPR 2025, aims to explore means to overcome current limitations in fully automated methods in long-form video understanding, by focusing on human-machine teaming approaches. While fully automated approaches have been the main focus of the research community, notable breakthroughs have been made in interactive video search and exploration as well. This workshop aims to showcase the main advantages and limitations of both paradigms. IViSE is structured in a challenge format with two tracks: Video Known-Item Search and Video Question Answering. These tasks can either be solved in a fully automatic fashion or in an interactive way where humans and machines collaborate. The interactive challenge is held as a live session at the workshop itself. This paper provides an overview of the participating teams, their approaches for both task types and the ways of approaching them. It presents the results for the fully automatic track and outlines the methods to be used in the live evaluation event held during the workshop in June of 2025.

1. Introduction

Holistic video understanding has long been a topic of interest in computer vision. Although the field has made tremendous progress in recent years, current state-of-the-art methods are often limited to video sequences with a total length measured in tens of seconds. Considering that narrative videos have an expected duration of minutes to hours, these methods are still not capable of covering a broad range of use cases. While machines are not very effective in dealing with such long-form content effectively yet, a collaborative team of humans and machines can be well supported by software applications in this endeavor.

The aim of the 1st international Workshop on Interactive Video Search and Exploration is to explore means to overcome current limitations in fully automated methods by focusing on human-machine teaming approaches for long-form video understanding. The workshop provides a venue to compare fully automated end-to-end approaches for video understanding and approaches in which humans and machines collaborate. The workshop is centered around a challenge on text-based video retrieval and visual question answering in a large collection of long videos from the first shard of the Vimeo Creative Commons Collection (V3C1) [8], which have an individual duration of between 3 minutes and 30 minutes per video, with 1000 hours of combined content. The challenge is split into two tracks: the *fully-automated track*, where queries are made available to participants beforehand and which they have to solve automatically without direct human intervention, and the *interactive track*, where queries are made available to participants during the workshop only and they have to solve them interactively in a human-machine team under a strict time

Table 1. Challenge participants at different tasks and tracks.

	Automatic		Interactive	
Team	KIS	QA	KIS	QA
WHU-NERCMS [5]	✓	✓		
nii-uit [4]	✓	✓	✓	✓
vifi [7]	✓	✓		
GenAI4E-TheBreaker [6]			✓	✓
GenAI4E-GameChanger [12]			✓	✓
certh_iti [3]	✓	✓		
diveXplore [10]			✓	✓
Exquisitor [11]	✓	✓	✓	✓

limit of five minutes per topic. This challenge format builds upon established challenges such as TRECVID [1] and DVU [2] on the fully-automatic side, and the Video Browser Showdown [14] and the Lifelog Search Challenge [13] on the interactive side.

2. Challenge Setup

The first instance of the IViSE challenge attracted 8 participating teams. Table 1 lists them, together with the tracks and tasks they participate in. Compared to the 17 Teams who participated in VBS 2025 [9], having results from five anticipated teams in the interactive challenge does not offer as much explanatory power. Nevertheless, having a direct comparison between automated and interactive approaches has the potential to offer novel insights, even with smaller sample sizes.

The challenge offers two types of tasks, Known-Item Search and Question Answering, which can both be solved in a fully automated way or interactively, with a human in the loop. For Known-Item Search queries, a textual description is provided which uniquely describes a segment of one particular video from the challenge dataset of 7475 videos. Queries of the Question Answering task are composed of a two-part textual description. The first part contains information on certain salient properties that are sufficient to uniquely identify a video. The second part then asks a question about that video. The question is generally completely unrelated to the first part of the description. The queries for the fully automated track are listed in Appendix A, and their corresponding ground-truth answers can be found in Appendix B.

3. Results of the Fully-Automated Track

In this section, we present the results of the fully-automated track for both KIS and QA tasks. Each team was allowed to submit up to five runs per task type, which were evaluated independently.

Table 2. Results of the Known-Item Search Tasks. Each task is scored on a scale from 0 to 1 and the final score is the sum of the 10 task scores. #V signifies the number of tasks for which the correct video was part of the result set and #S signifies the number of tasks where the correct video segment was part of the result set.

Team	Run	Score	#S	#V
nii-uit	1	9.9	10	10
vifi	4	9.8	10	10
vifi	5	9.8	10	10
vifi	3	9.6	10	10
nii-uit	2	8.9	10	10
vifi	1	8.9	9	9
Exquisitor	1	8.3	9	9
Exquisitor	2	7.9	8	8
Exquisitor	4	7.8	8	8
vifi	2	7.6	8	8
Exquisitor	3	7.3	8	8
WHU-NERCMS	3	6.0	6	9
WHU-NERCMS	2	5.8	6	9
WHU-NERCMS	1	4.0	4	9
nii-uit	3	3.5	5	8
certh_iti	1	3.5	4	8
certh_iti	2	1.6	3	4

3.1. Known-Item Search

Five teams submitted a total of 17 runs for the known-item search task. The results are listed in Table 2 and the rank distribution is visualized in Figure 1.

The scores of this task were remarkably high. Two of the participating teams – nii-uit and vifi – managed to find the correct video segments for all ten tasks and even had them ranked first in most cases. All teams and most runs were able to correctly identify a vast majority of the target videos, but some teams struggled with correctly identifying the segment boundaries. This can be seen in the difference between the #S and #V columns in Table 2, showing the correctly identified segments and videos, respectively.

There appears to be a difference in difficulty across the tasks, as shown in Table 3. Task IVISE25-KIS-A03 appears to have been the most difficult, only being solved correctly in half of the submitted runs, while task IVISE25-KIS-A10 was the easiest, being correctly solved in all runs and almost always on rank 1, resulting in an almost perfect score.

3.2. Question Answering

For the Question Answering task, we received 16 runs from five teams. The results are listed in Table 4 and the ranks are illustrated in Figure 2. The scores are generally lower in this task compared to the known-item search task, resulting from the added complexity of not only identifying the correct video from the collection but extracting sufficient se-

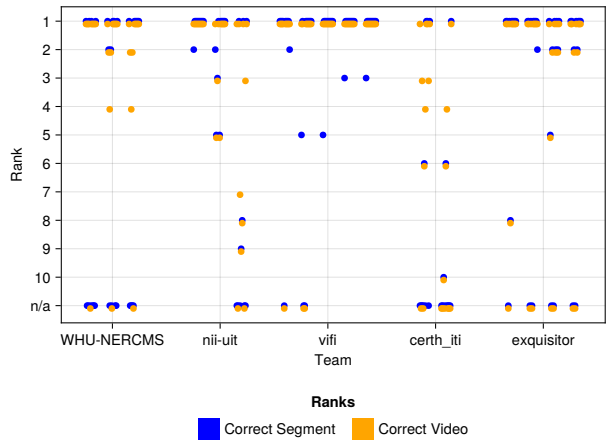

Figure 1. Rank distribution of Known-Item Search tasks

Table 3. Aggregated performance per known-item search task.

Task	Mean Score	#Solved	#Video
IVISE25-KIS-A01	0.741	14	16
IVISE25-KIS-A02	0.653	12	14
IVISE25-KIS-A03	0.406	8	12
IVISE25-KIS-A04	0.688	12	14
IVISE25-KIS-A05	0.418	10	10
IVISE25-KIS-A06	0.841	15	17
IVISE25-KIS-A07	0.871	15	16
IVISE25-KIS-A08	0.765	13	15
IVISE25-KIS-A09	0.700	12	16
IVISE25-KIS-A10	0.988	17	17

mantics to correctly formulate an answer to the posed question. While no team managed to correctly answer all questions, WHU-NERCMS, vifi, and nii-uit correctly answered 7 out of 10 questions. In the case of WHU-NERCMS, all correct answers were on the first rank of their submissions, resulting in the maximum score. All three these teams also correctly identified all target videos, but were then unable to answer some of the questions.

When considering Table 5, which shows the results aggregated per task, we can see clear differences in difficulty between the different tasks. Three questions that remain consistently unanswered by any team: IVISE25-QA-A01, IVISE25-QA-A07, and IVISE25-QA-A09. In the case of IVISE25-QA-A01, all teams correctly identified the target video. They were then, however, unable to correctly identify the name of the call sign of the radio station. This despite some methods having identified the parts of the video where it is shown or mentioned, and in some cases even provided a partial transcript that mentions the correct call sign. In all cases, they did, however, not draw the correct conclusion, resulting in an incorrect answer. IVISE25-QA-

Table 4. Results of the Question Answering Tasks. Each task is scored on a scale from 0 to 1 and the final score is the sum of the 10 task scores. #V signifies for how many tasks the correct video was part of the result set and #A signifies for how many tasks the correct answer was part of the result set.

Team	Run	Score	#A	#V
WHU-NERCMS	1	7.0	7	10
WHU-NERCMS	2	7.0	7	10
vifi	1	6.3	7	10
vifi	3	6.3	7	10
nii-uit	3	6.5	7	10
nii-uit	1	5.8	6	10
nii-uit	2	5.8	6	10
nii-uit	4	5.8	6	10
vifi	2	5.2	6	9
Exquisitor	2	4.2	5	7
Exquisitor	3	3.6	4	6
Exquisitor	4	3.6	4	6
Exquisitor	1	3.4	4	6
certh_iti	1	2.4	3	6
certh_iti	2	2.4	3	5
certh_iti	3	1.8	3	6

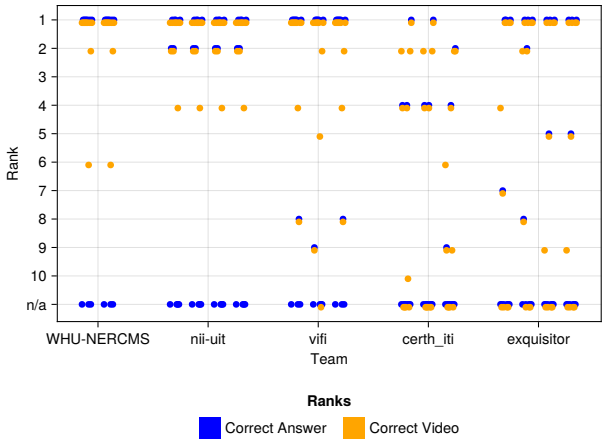

Figure 2. Rank distribution of Question Answering tasks

A07 is an interesting failure case, as it did not only require the teams to identify a specific flag that was small and only briefly visible, but also have sufficient world knowledge to correctly identify the flag. While some teams succeeded in the former and correctly localized the flag in question and even mentioned its colors, no team was able to map this information correctly to the flag's origin. The case of IVISE25-QA-A09 is also interesting, since it did not target any specific part of the video but asked about a property of the video as a whole. While all teams correctly identified the target video, all methods also substantially underre-

ported the number of distinct aerial shots.

Table 5. Aggregated performance per question answering task.

Task	Mean Score	#Solved	#Video
IVISE25-QA-A01	0.000	0	16
IVISE25-QA-A02	0.577	13	13
IVISE25-QA-A03	0.723	13	13
IVISE25-QA-A04	0.431	9	9
IVISE25-QA-A05	0.700	13	13
IVISE25-QA-A06	0.931	16	16
IVISE25-QA-A07	0.000	0	11
IVISE25-QA-A08	0.154	4	8
IVISE25-QA-A09	0.000	0	16
IVISE25-QA-A10	0.992	16	16

4. Preview of the Interactive Track

Since the interactive track is evaluated live during the workshop held at CVPR 2025, no results can be reported at the time of writing.

5. Conclusion

This paper presented an overview of the 1st International Workshop on Interactive Video Search and Exploration, as well as its associated challenge. The challenge managed to attract 8 teams who participated in at least one of the two task types, either interactively or in a fully-automated fashion. The results from the fully-automated track indicate that there has been substantial progress in recent years in both video search and video question answering and that these methods can be combined into fully-automated video retrieval and question answering pipelines. While the results were remarkably good, the challenge still revealed some aspects that current methods are not able to correctly handle. Future instances of the challenge will hence increase the difficulty of the tasks, with a greater focus on addressing those challenging aspects. The interactive track is evaluated at the workshop itself and no results are available at the time of writing. Once the results are available, we will study and contrast them with the ones achieved by the fully-automatic approaches in a follow-up analysis.

References

[1] George Awad, Jonathan Fiscus, Afzal Godil, Lukas Diduch, Yvette Graham, and Georges Quénot. Trecvid 2024 - evaluating video search, captioning, and activity recognition. In *Proceedings of TRECVID 2024*. NIST, USA, 2024. 2

[2] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. Hlvu: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 355–361, 2020. 2

[3] Damianos Galanopoulos, Andreas Goulas, Antonios Leventakis, Ioannis Patras, and Vasileios Mezaris. An LLM Framework for Long-form Video Retrieval and Audio-Visual Question Answering Using Qwen2/2.5. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[4] Bao Tran Gia, Khiem Le, Tien Do, Dung Mai Tien, Thanh Duc Ngo, Duy-Dinh Le, and Shin'ichi Satoh. VRAG: Retrieval-Augmented Video Question Answering for Long-Form Videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[5] Heng Liu, Siru Jiang, Fangyun Duan, Yongzhe Lyu, Xiusong Wang, Hanlin Ge, and Chao Liang. CadenceRAG: Context-Aware and Dependency-Enhanced Retrieval Augmented Generation for Holistic Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[6] Tinh-Anh Nguyen-Nhu, Huu-Loc Tran, Nguyen-Khang Le, Minh-Nhat Nguyen, Tien-Huy Nguyen, Long Hoang Huu Nguyen, Huu-Phong Phan-Nguyen, Huy-Thach Pham, Quan Nguyen, Hoang M. Le, and Vinh Quang Dinh. A Lightweight Moment Retrieval System with Global Re-Ranking and Robust Adaptive Bidirectional Temporal Search. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[7] Khanh-An C. Quan, Qui Ngoc Nguyen, and Duc-Tuan Luu. Toward Automation in Text-based Video Retrieval with LLM Assistance. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[8] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. V3C - A research video collection. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I*, pages 349–360. Springer, 2019. 1

[9] Luca Rossetto, Klaus Schoeffmann, Cathal Gurrin, Jakub Lokoč, and Werner Bailer. Results of the 2025 video browser showdown. arXiv, 2025. 2

[10] Klaus Schoeffmann and Mario Leopold. AI-based Video Content Understanding for Automatic and Interactive Multimedia Retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[11] Ujjwal Sharma, Omar Shahbaz Khan, Stevan Rudinac, and Björn Þór Jónsson. Can Relevance Feedback, Conversational Search and Foundation Models Work Together for Interactive Video Search and Exploration? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[12] Huu-Loc Tran, Tinh-Anh Nguyen-Nhu, Tien-Huy Nguyen, Nhat Minh Nguyen Dich, Anh Dao, Duc Huy Do, Huu-Phong Phan-Nguyen, Quan Nguyen, Hoang M. Le, and Vinh Quang Dinh. Towards Efficient and Robust Moment Retrieval System: A Unified Framework for Multi-Granularity Models and Temporal Reranking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025. 2

[13] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoc, Ladislav Peska, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, Björn Þór Jónsson, Luca Rossetto, An-Zi Yen, Ahmed Alateeq, Naushad Alam, Minh-Triet Tran, Graham Healy, Klaus Schoeffmann, and Cathal Gurrin. Comparing interactive retrieval approaches at the lifelog search challenge 2021. *IEEE Access*, 11:30982–30995, 2023. 2

[14] Lucia Vadicamo, Rahel Arnold, Werner Bailer, Fabio Carrara, Cathal Gurrin, Nico Hezel, Xinghan Li, Jakub Lokoc, Sebastian Lubos, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peska, Luca Rossetto, Loris Sauter, Klaus Schöffmann, Florian Spiess, Minh-Triet Tran, and Stefanos Vrochidis. Evaluating performance and trends in interactive video retrieval: Insights from the 12th VBS competition. *IEEE Access*, 12:79342–79366, 2024. 2

A. Queries used for the Fully Automated Track

IVISE25-KIS-A01 Close-up of the hand of an elderly Asian man putting down a green beer bottle on the table. There are a full beer glass, some bowls and a box containing a cognac bottle on the table. The camera moves to the man's face, before going back to his hands, as he uses chopsticks to put food from one bowl into sauce in another bowl.

IVISE25-KIS-A02 A shot of mostly wooden outdoor fitness equipment, installed between trees. The camera pans to the left, before the next shot shows a closeup of a wooden sculpture, with the words "Peace, Hope & Love" engraved.

IVISE25-KIS-A03 A sequence of shots taken from a moving motorbike. In the first shot we see a view under the rider's left arm, the handlebar, both mirrors and the rider's hands are visible. The motorbike moves along a country road, and a red triangular street sign indicates a railway crossing coming up. The following shot is a view down to the road on the right side of the motorbike, first showing the right part of the handlebar and then moving forward.

IVISE25-KIS-A04 A woman in a wedding dress is standing with outstretched arms on an embankment made of rocks. She is holding a white translucent shawl or vail behind her. In the background, there is a cliff face with a waterfall feeding into a little lake. The next shot shows her more zoomed in as she is wrapping her shawl/vail around her.

IVISE25-KIS-A05 Two shots from the window of a moving train: In the first one we see some of the cars in front on the right, and a steep wall, covered with plants, close to the track on the left. In the second shot the view is out of a window on the other side, with the train on the left and a young man standing at the door, and bushes on the right. A black and white mile marker with 101 is visible.

IVISE25-KIS-A06 A man wearing red gloves is standing on ice, only his lower half is in frame. At his feet are several fish, still alive. He bends down and throws some of the fish back into the water through a hole in the ice. He puts the remaining fish into a black sled next to him. He picks up a blue rope and pulls away the sled.

IVISE25-KIS-A07 Several shots of old electronics. The first one shows a wall of stacked desktop PCs, the next a close-up of a processor socketed in a mainboard. The next shots respectively show a closeup of a hand holding a processor pins-up, a box full of old processors, and a box full of old mobile phones.

IVISE25-KIS-A08 Two dogs in a stream at the base of a waterfall. One dog has light yellow fur, the other has black fur and a blue collar. After getting out of the water, they walk up an incline next to the waterfalls, accompanying some hikers.

IVISE25-KIS-A09 A man is sitting in a camping chair outside, drinking from a brown glass bottle. He is holding a green bottle in his other hand. Next to him is another chair with a large white teddy bear in it. The man turns to the bear and places the green bottle in the bear's arm.

IVISE25-KIS-A10 A timelapse of several people mounting large posters with several color images and black text onto wooden structures. The structures are curved so they stand upright to present the posters, but they are laid flat on the floor for mounting them. The whole sequence is filmed from a fixed camera placed high in a corner of the room.

IVISE25-QA-A01 Grainy black/white shot of a radio antenna mast, panning down until the roof of a car becomes visible, followed by a shot showing a hand turning both dials of a radio receiver. What was the call sign of the first radio station founded by the man portrayed in this documentary?

IVISE25-QA-A02 The video is filmed from the perspective of a paraglider flying over a lush green landscape. There is a deep blue lake surrounded by grass-covered hills, along which the paraglider is flying. In the distance, the ocean is visible. The paraglider is often flying very close to the steep hills on their right. At the very beginning of that video, there is a title card indicating the year and the location the video was recorded. Where and when was that?

IVISE25-QA-A03 A group of people is holding a workshop in a room with pastel green walls with white trim, and parquet flooring. Throughout the video, the people are shown in various constellations in the room, intercut with several short sequences where individual people are talking directly into the camera. There are some shots with different arrangements of chairs and tables. On one side of the room, there is a raised stage with a short red curtain at the top. What institution hosted that workshop?

IVISE25-QA-A04 An urban setting filmed in black and white. Early in the video, there is a shot where the camera is looking out through the window of a moving train. The video is then composed of several outdoor shots with falling snow. During the whole video, there is music playing. According to the end card, what is the name of the song?

IVISE25-QA-A05 Two shots showing a wooden hut with a chimney made of stone on the left end. It is sunny, but the ground and the roof of the hut are mostly covered with snow. A woman (wearing jeans and greenish/petrol jacket, gloves, bonnet) and a child (jeans, blue jacket, gloves, bonnet) walk towards the hut. In the second shot, the woman opens the door of the hut, which makes a squeaking sound. Which drink does the woman serve when they are back in their car?

IVISE25-QA-A06 A woman is sitting in a chair, playing the accordion and singing a song. She has long light-brown hair and wears a red sweater. She is alternatingly shown singing from several perspectives for several minutes. What language is the song in?

IVISE25-QA-A07 Two shots of a wooden pavilion on a public square, with benches below and cars parked in front. A long red boat is stored below the ceiling. The first shot shows a closeup of the boat, the second a larger view of the square. This clip was made on a sailing trip. Which is the lowest flag on the rope down from the mast to the right side of the bow of the sailboat?

IVISE25-QA-A08 Two women are having a conversation over the telephone. One is trying to hang pictures from the wall and is asking the other for advice. The video cuts back and forth between them. The woman with the pictures then drives to a hardware store. What are the names of the two women?

IVISE25-QA-A09 A series of aerial shots of buildings in different locations. The shots include a round white building with a sign saying "CAPITOL RECORDS". Another shot shows a view of the "GRAND LAKE THEATRE". There is a watermark in the corner of the entire video. How many different aerial video sequences are in the video in total?

IVISE25-QA-A10 A man wearing dark trousers and jacket, a scarf and a horse mask walks down a path. There is a wooden bench on the right, cars and houses in the background, and litter on the path and grass. He shakes his head and puts his hand to the head. The camera follows him as he walks to another wooden bench and sits down. Which writer is quoted at the start of this video?

B. Ground truth answers for the Fully Automated Track

IVISE25-KIS-A01 video 00726, seconds 296 to 316

IVISE25-KIS-A02 video 04873, seconds 124 to 132

IVISE25-KIS-A03 video 05739, seconds 203 to 205

IVISE25-KIS-A04 video 04755, seconds 53 to 62

IVISE25-KIS-A05 video 01596, seconds 71 to 80

IVISE25-KIS-A06 video 06884, seconds 36 to 92

IVISE25-KIS-A07 video 01425, seconds 88 to 105

IVISE25-KIS-A08 video 05506, seconds 117 to 157

IVISE25-KIS-A09 video 07309, seconds 201 to 206

IVISE25-KIS-A10 video 04962, seconds 27 to 44

IVISE25-QA-A01 "WW2A" (video 05052)

IVISE25-QA-A02 "Azoren 2016" (video 06208)

IVISE25-QA-A03 "The University of Edinburgh" (video 04222)

IVISE25-QA-A04 "Talking To Ghost" (video 03343)

IVISE25-QA-A05 "Coffee" (video 02431)

IVISE25-QA-A06 "French" (video 01155)

IVISE25-QA-A07 "Basque, Euskadi" (video 00063)

IVISE25-QA-A08 "Tori and Marsha" (video 03288)

IVISE25-QA-A09 "40" (video 04093)

IVISE25-QA-A10 "Balzac" (video 03087)